DOCUMENT RESUME

ED 235 225 TM 830 646

AUTHOR Hambleton, Ronald K.
TITLE Standard-Setting: State of the Art, and Future Prospectus. Report No. 142.
INSTITUTION Massachusetts Univ., Amherst. Laboratory of Psychometric and Evaluative Research.
PUB DATE 83
NOTE 20p.
PUB TYPE Viewpoints (120)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Academic Standards; Decision Making; *Evaluation Methods; Evaluation Needs; Psychometrics; *Research Methodology; *Research Problems
IDENTIFIERS *Standard Setting

ABSTRACT

This paper offers answers to nine important questions concerning standard-setting issues and methods: (1) Should normative or content-referenced standards be used? (2) Different standard-setting methods yield different results. What is your reaction to this finding? That of your clients? Does this finding present a problem for the application of various standard-setting methods? (3) Assess the adequacy of the grounding of various methods of standard-setting in psychological and/or psychometric theory. (4) Should standards be validated? If so, how can this be done? (5) Within the context of the overall problem of standard-setting, consider the roles that are or should be played by the client, technical consultant, candidates, and other actors. What is the proper role of the public? (6) To what extent should standard-setting processes attempt to formally incorporate social/political considerations into the decision-making process? (7) What are the ethical responsibilities of the technical consultant? (8) Why have developments come so slowly? How do you view the future of standard-setting? (9) What are the key short-term and long-term research problems that should be addressed? (Author)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED235225

Standard-Setting: State of the Art, and Future Prospectus

Ronald K. Hambleton
University of Massachusetts, Amherst

Abstract

Answers to nine important questions concerning standard-setting issues and methods are offered in the paper:

1. Should normative or content-referenced standards be used?

2. Different standard-setting methods yield different results. What is your reaction to this finding? That of your clients? Does this finding present a problem for the application of various standard-setting methods?

3. Assess the adequacy of the grounding of various methods of standard-setting in psychological and/or psychometric theory.

4. Should standards be validated? If so, how can this be done?

5. Within the context of the overall problem of standard-setting, consider the roles that are or should be played by the client, technical consultant, candidates, and other actors. What is the proper role of the public?

6. To what extent should standard-setting processes attempt to formally incorporate social/political considerations into the decision-making process?

7. What are the ethical responsibilities of the technical consultant?

8. Why have developments come so slowly? How do you view the future of standard-setting?

9. What are the key short-term and long-term research problems that should be addressed?

TM 830 654

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☒ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

R. Hambleton

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Standard-Setting: State of the Art, and Future Prospectus[1]

Ronald K. Hambleton
University of Massachusetts, Amherst

In the time I have this afternoon I would like to offer some brief answers to the nine important questions (listed in Appendix A) posed by John Meskauskas, the organizer of this symposium. I will not give equal time to the nine questions in my remarks this afternoon since I have limited expertise and/or no informed opinions in several of the areas.

1. Normative Versus Content-Referenced Standards

The use of normative standards (i.e., setting the pass rate without regard for the examinee population or the difficulty of the test) is clearly inconsistent with the basic goal of criterion-referenced testing. That goal is to assess each candidate in relation to a set of clearly stated objectives or tasks. The score a candidate receives on a CRT or the mastery status he/she is assigned to should not depend upon the performance of other candidates taking the test. On the other hand, it would be foolish to implement a standard using one of the content-referenced methods without any knowledge of the implications. Information about actual (or expected) score distributions should be provided to standard-setters. Both Richard Jaeger in North Carolina (Jaeger, 1982) and Paul Williams in Maryland have been very successful with

---

[1] Laboratory of Psychometric and Evaluative Research Report No. 142. Amherst, MA: School of Education, University of Massachusetts, 1983.

such a strategy. In fact, I would go one step further. I would make available to groups setting a standard, information such as item difficulty and the percentage of candidates selecting each answer choice. There is ample evidence in the psychometric literature revealing the many seemingly minor aspects of items that influence item difficulty. The judgmental review will be more informed when reviewers have access to both score distributions and the complete item analysis information.

In summary, I certainly don't want to defend the use of normative standards, but I do support the setting of standards influenced by the use of test score distributions from appropriate groups of examinees.

2. Different Standard-Setting Methods--Different Results

I have never worried about this question nor felt it was worthy of all the attention it has received from researchers. I feel this way for three reasons: First, there is no reason to expect various methods to lead to the same standard. These methods are often based on different notions of minimal competence and utilize substantially different kinds of information (e.g., judges ratings of test items versus distributions of examinee test scores). For example, statistical tests used by educational researchers do not always lead to the same conclusions but there is no cry that I hear for discontinuing the use of statistical methods in research studies. Second, the generalizability of results from comparative studies is often very limited because the effects of any changes in the implementation of a method are typically unknown. For example, how much difference in the selected standard would there be if judges were provided with the correct answers to the test questions they reviewed? Finally, I feel that the resulting standard is not the most appropriate consideration

in choosing a method. That is, a method should not usually be chosen specifically because it leads to higher or lower standards than standards obtained from other methods. For example, some groups have preferred the Nedelsky method (Nedelsky, 1954) because it usually leads to a lower standard than other content standard-setting methods. The field of competency testing would be substantially better served if more attention from researchers, and standard setters, were given to the total standard-setting process which includes everything from the initial discussions of the importance of the decisions to be made with the test and available resources, to deciding on the scope and specifics of the procedure (for example, will judges be used, and if so, how will they be selected? Will test results be used, and if so, how will they be used? What will the nature be of the judging task -- for example, evaluating distractors or test items?), to the implementation of the procedure, and finally, to combining the various pieces of data in a specified way to arrive at a final standard (or standards).

3. Adequacy of Grounding of Standard-Setting Methods in Psychometric and Psychological Theory

I can't think of any psychological theory providing a sound basis for standard-setting except on a few criterion-referenced tests which provide specific information about child developmental levels. Here, there may be psychological theories and research results that relate developmental levels to specific levels of performance on (say) learning tasks. Such theories may provide a basis for standard-setting. For example, if a child has not reached a particular stage in (say) Piaget's paradigm for conservation of numbers, then there is no good reason for exposing a child to a high level skill in mathematics.

With respect to psychometric foundations, I only note here that there now exists a well-developed criterion-referenced test theory. The present psychometric theory provides a framework for assessing the reliability and validity of any standard that is set. In that sense at least I think present standard-setting methods are "well-grounded" in psychometric theory.

4. Necessity of Validating Standards

Should it be done? Of course. To ask the question, "Is the standard valid?" is to ask, "Will the particular choice of a standard result in a valid classification of examinees?" The question is approached in the same way as any other test validity question. Threats to validity should be addressed and studied. Evidence must be compiled to support or refute the mastery classifications and intended uses of the classifications. Validity evidence usually comes in two broad forms: evidence pertaining (1) to the selection of a standard, and (2) to the actual functioning of the standard.

The first type of evidence comes in the form of a measure of agreement reflecting the similarity of standards in different groups (e.g., national leaders, university instructors, and practicing professionals). With respect to the latter type of evidence, information on the percent of examinees who are correctly classified in contrasting groups or the percent of examinees who are classified in the same way by the test of interest and an external criterion measure would be especially useful. Other experiments can be easily thought of. The available time and money, and importance of the test, will help to determine the emphasis.

With respect to reliability information, of interest would be the distribution of standards for judges from similar backgrounds. This type of information would be especially meaningful when each judge's ratings are prepared independently. Alternately, when comparable groups are formed, the distribution of standards across groups will be of interest or some other appropriate measure of consistency or agreement.

5. Roles of Various Players in the Process (Client, Technical Consultant, Public)

Of course the answer to this difficult question will be situation specific. The technical consultant's role is the easiest to define. His/her task is to insure that (1) the standard-setting plan is thoughtfully developed with appropriate input from relevant groups and individuals; (2) the plan is fully implemented and when revisions are made, justifications are sound; and (3) the total process is documented with reasons available for all of the key decisions; and (4) the plan and its implementation are consistent with up-to-date standard-setting methods and procedures. The acceptability of the resulting standard will depend greatly on the appropriateness and justification of the process. It is likely to depend very little on which particular method was implemented Figure 1 provides some guidance to the consultant on areas which must be addressed in the planning stages for a standard-setting study.

6. Incorporation of Social/Political Considerations

For this question I will limit my discussion to political matters associated with using decision theory to set cut-off scores. The choice of cut-off score, for example, in the context of basic skills programs and

certification exams, will ultimately impact on the perceptions the public has of these testing programs. What will these perceptions be when policy-makers announce that a group of carefully selected judges, in some cases, 100s, set a standard of 70% and the optimal cut-off score as identified with decision theory methodology was 60%, and so the latter value will be used to make mastery/non-mastery decisions? What will the judges who labored over their task think? Will the public feel that the cut-off score was lowered so that more persons could pass and make the institutions training the examinees appear to be better than they are? And, if the optimal cut-off score is set at 80% will every candidate between 70% and 79% sue? And, will candidates scoring below 70% sue on the grounds that the resulting confusion over standards and cut-off scores could only be created by fools and so why should they be penalized for the incompetence of the test designers?

But, there is more. It is common at the district and state levels in the U.S. to compare student performance in different subject areas. While admittedly faulty reasoning, policy makers often feel they can be fair to all concerned by establishing common standards across subject areas. The use of optimal cut-off scores resulting (in general) in different cut-off scores across subject areas would lead to heated debates. For example, the mathematics people could claim their standards were higher, and thereby explain away the fact that more students failed mathematics than other subjects. They might take pride in the results, or they might even feel that they and their students are being unfairly treated. Clearly, in important testing situations it will be difficult to predict all of the questions which might arise when standards and cut-off scores differ, and even more difficult to respond to the questions!

7. Ethical Responsibilities of the Technical Consultant

The ethical responsibilities for a technical consultant are no different in the context of standard-setting problems than the ethical responsibilities of a measurement specialist at any time he/she is working with a client. The measurement specialist is responsible for providing accurate technical information and exhibiting sound professional judgment in using measurement models and procedures. No more can be expected and certainly no less.

8. Slowness of Developments/Future of Standard-Setting

First of all I disagree with the premise of the first part of the question. It seems to me that considerable progress has been made since the early 1970's. Before 1971, probably the only standard-setting methods in the psychometric literature were the Nedelsky and the contrasting groups methods. In 1983 it can be said that there is no shortage of (1) discussions of issues associated with standard-setting (for example, see the special Winter 1978 issue of the Journal of Educational Measurement), (2) methods for standard setting (Meskauskas, 1976; Millman, 1973; Popham, 1981), and (3) reviews of these methods (Glass, 1978; Shepard, 1980a, 1980b). Also, Hambleton (1978), Linn (1978), and Shepard (1976) have offered recommendations for setting standards, and Popham (1978) and Livingston and Zieky (1982) have offered steps for implementing several of the more promising methods. Perhaps the pace of new research findings has lagged behind the pace with which competency testing programs have been implemented in the schools and at the state level, and the implementation of certification and licensure exams. Still, the pace seemed very fast to me.

Consider norm-referenced testing. Norm-referenced testing models and methods have been under development for seventy years and the work is still going strong! Consider, for example, the present amount of work on item response theory (IRT) and generalizability theory.

9. Research Topics[1]

There are many research questions that need to be addressed in the coming years so that factors influencing the results can be identified and their effects estimated. Figure 2 provides a long list of important research questions. Time permits a brief consideration of only two. Most of the questions are discussed by Hambleton and Powell (1983).

The first question is:

> When judges are arranged into working groups, what is the optimal group size, and should the groups be formed homogeneously or heterogeneously?

Of course, the first question is whether or not to form working groups. Working groups should be formed when (1) there is interest in promoting discussion among the participants and the total group of standard-setters is too large to permit effective discussion, (2) comparability of standards across similar groups, or, if you like, the reliability of a standard, is of interest (when there are dissimilar factions present they should be distributed uniformly across the groups), or (3) comparability of standards, or, if you like, the validity of a standard across dissimilar groups (e.g., national leaders in the field, versus practicing

---

[1]Most of the material in this section is from a paper by Hambleton and Powell (1983).

professionals), is of interest. The second situation calls for groups that can be considered to be very similar, but within each group there will be substantial heterogeneity to reflect the diversity among individuals involved in the standard-setting procedure. Of special interest is the stability of the selected standard across similar groups. The third situation calls for groups that represent the various factions selected to participate in the standard-setting procedure. The groups will often differ substantially in backgrounds, perspectives and priorities. However, there is often relatively more similarity among members of each group. Of interest in this situation is the similarity of the obtained standard across different groups. Since both approaches for grouping judges have considerable merit, incorporating both within a single standard-setting procedure would seem to be highly desirable.

The second question is:

> Should item content be considered with or without examinee item performance data?

In reviewing item content, some standard-setters have advocated that judges consider the performance levels (item difficulties) of relevant examinee samples on the test items of interest. When item performance data are used, it is essential that they be obtained on groups of examinees who can be clearly described. This type of data, however, is usually both time-consuming and expensive to collect. Still, most researchers in the field highly endorse the desirability of this type of information (for example, see Linn, 1978; Popham, 1981; Shepard, 1980b). Popham (1981), for example, recommends that the test results of non-instructed, recently

instructed, and previously instructed groups be made available to standard-setters.

The difficulty of making absolute judgments about item content (that is, judgment in the absence of any performance data) is well-known. At the same time, in the context of criterion-referenced testing it is not usually considered desirable to weight normative data too heavily either in test construction or in standard setting. One possible compromise between the absolute and normative positions could involve supplying extensive item analysis data to judges who would use the data to explore the cognitive processes involved (for the examinee population) in answering each item. Item analysis data might include among other things the percent of high- and low-scoring examinees choosing each item alternative, along with item difficulty and discrimination statistics. The standards thus set would be "absolute" but informed by normative data.

One of the major sources of difficulty in setting standards is that we do not usually know exactly what our test items are measuring; in the ideal situation we know what it is we intend them to be measuring, but that is not usually perfectly operationalized in actual items. If we were completely knowledgeable about the relation of responses to an item and the objective we wish to measure, we would probably feel more confident in setting standards on the basis of actual items. This is probably the source of our desire to see performance data when judging items: we want some evidence that the item is really measuring what we intend. However, we should be very clear about the fact that what we want here is construct validity evidence for the test items. This suggests that we should be looking at a wide variety of data on the item, and not simply at p-values.

Extensive item analysis data would be helpful in this endeavor, as would experimental data possibly collected expressly for a standard-setting study. The general point here is that it is not possible to make absolute judgments about performance on test items without knowing exactly what that performance reflects.

Conclusion

I am pleased with the progress that has been made in the last 10 to 15 years. Issues have been identified, new methods have been developed and field-tested; and guidelines for implementing many of the methods are available.

My major concern with the work going on is that too often groups setting standards define the task too narrowly. Standard-setting is far more than choosing and implementing a method. It is essential to consider the total standard setting process which includes the planning, implementation, data analysis, and documentation stages. A rationale and justification for each decision made in the standard-setting process should be available and documented. In addition, the process should be carried out according to the plan and variations should be documented.

Finally, when sound professional judgment is combined with state-of-the-art standard-setting methods and procedures, districts, states, and organizations can be very proud of their efforts. Nothing more can be expected. They may still be sued, but at least they will have a strong position from which to defend themselves.

References

Glass, G. V. Standards and criteria. Journal of Educational Measurement, 1978, 15, 277-290.

Hambleton, R. K. On the use of cut-off scores with criterion-referenced tests in instructional settings. Journal of Educational Measurement, 1978, 15, 277-290.

Hambleton, R. K., & Powell, S. A framework for viewing the process of standard-setting. Evaluation and the Health Professions, 1983, in press.

Jaeger, R. M. An iterative structured judgment process for establishing standards on competency tests: Theory and application. Educational Evaluation and Policy Analysis, 1982, 6, 461-475.

Linn, R. L. Demands, cautions, and suggestions for setting standards. Journal of Educational Measurement, 1978, 15, 301-308.

Livingston, S. A., & Zieky, M. J. Passing scores: A manual for setting standards of performance on educational and occupational tests. Princeton, NJ: Educational Testing Service, 1982.

Meskauskas, J. A. Evaluation models for criterion-referenced testing: Views regarding mastery and standard-setting. Review of Educational Research, 1976, 46, 133-158.

Millman, J. Passing scores and test lengths for domain-referenced measures. Review of Educational Research, 1973, 43, 205-216.

Nedelsky, L. Absolute grading standards for objective tests. Educational and Psychological Measurement, 1954, 14, 3-19.

Popham, W. J. Setting Performance Standards. Los Angeles: Instructional Objectives Exchange, 1978.

Popham, W. J. Modern Educational Measurement. Englewood Cliffs, NJ: Prentice-Hall, 1981.

Shepard, L. A. Setting standards and living with them. Florida Journal of Educational Research, 1976, 18, 23-32.

Shepard, L. A. Technical issues in minimum competency testing. In D. C. Berliner (Ed), Review of Research in Education (Vol. 8). Itasca, IL: F. E. Peacock Publishers, 1980. (a)

Shepard, L. A. Standard setting issues and methods. Applied Psychological Measurement, 1980, 4, 447-467. (b)

Appendix A

Questions Addressed By the Panel

1. Should normative or content-referenced standards be used?

2. Different standard-setting methods yield different results. What is your reaction to this finding? That of your clients? Does this finding present a problem for the application of various standard-setting methods?

3. Assess the adequacy of the grounding of various methods of standard-setting in psychological and/or psychometric theory.

4. Should standards be validated? If so, how can this be done?

5. Within the context of the overall problem of standard-setting, consider the roles that are or should be played by the client, technical consultant, candidates, and other actors. What is the proper role of the public?

6. To what extent should standard-setting processes attempt to formally incorporate social/political considerations into the decision-making process?

7. What are the ethical responsibilities of the technical consultant?

8. Why have developments come so slowly? How do you view the future of standard-setting?

9. What are the key short-term and long-term research problems that should be addressed?

Figure 1. A listing of context-setting variables in a standard-setting procedure.

Context-Setting Variables

1. Importance of the decisions being made.

   Consider these questions:

   a. How many individuals are directly and indirectly affected by the decisions to be based on the test? (For example, what resources are available to remediate those who are identified as non-masters? How many non-masters can be handled effectively?)

   b. What are the possible educational, psychological, financial, and other consequences of the decisions?

   c. What is the duration of the consequences?

2. Availability of resources.

   Consider these questions:

   a. How much money, time, and material are available to carry out the standard-setting procedure? What level of technical expertise is available (in-house or otherwise) to complete the work?

   b. Are the resources fixed or flexible? (For example, can the budget items for the total project be rearranged to hire additional consultants and collect additional data? Can additional time be obtained to carry out the procedure?)

3. Test format(s), content, and length.

   Consider these questions:

   a. What item formats are used in the test (e.g., multiple-choice, true-false, short-answer, essay, performance)?

   b. What content areas are covered by the test? (A test blueprint or item specifications would usually suffice to address the question.)

   c. How many test items are presented in each of the item formats?

4. Laws.

   Consider these questions:

   a. Do the laws state the groups who should be involved in the standard-setting procedure?

   b. Does the law provide a definition of the "minimally competent candidate"?

   c. Do the laws address whether "compensatory" or "non-compensatory" decision-making models should be used?

5. History.

   Consider these questions:

   a. Is the testing program relatively new?

   b. Will there be an opportunity later to review the standard(s)?

Figure 2. A listing of questions which must be addressed in a standard-setting procedure.

Judges

a. Which demographic variables should be used in selecting judges?

b. How should names of possible judges be generated?

c. Which individuals should be involved in the judge selection process (and why)?

d. How many judges should be selected to participate?

e. Should judges be volunteers or should they be conscripted?

f. Should judges be selected to be representative of some constituency?

g. Should "expert" judges be preferred over representatives of groups of interest?

h. When judges are arranged into working groups, what is the optimal group size, and should the groups be formed homogeneously or heterogeneously?

i. Should data from judges be discarded when there is reason to believe that they were unqualified to do the job, or carried out the task in a "sloppy" fashion? Should specific steps be taken to identify "poor" judges?

j. Should judges be paid for their time?

Judgments of Items

a. Should items be judged individually or "globally" in groups?

b. Should item distractors be judged individually or should the total item be judged?

c. Should all of the test items be judged or only a sample?

d. Should judgments be made of the actual test items or "example" items?

e. Should the judgments be based on a review of the test items or the objectives they were prepared to measure?

f. Should item content be considered with or without examinee item performance data?

g. Should items be sorted in any way to aid judges?

h. Should the correct answers be identified for judges?

i. Should judges prepare their ratings independently or as part of a group process in which each item is deliberated on and discussed before final judgments are made?

Nature of the Judgmental Process

a. Should a single judgment/item be made or should an iterative process be used where judges have an opportunity to revise their opinions based upon feedback in one form or another from other judges?

b. In combining judgments across individuals and/or groups should weights be used (1) to reflect the perceived importance of each individual and/or group, (2) group size, etc.?

c. Should differences between individuals and/or groups be resolved by averaging, by reaching consensus, or some other method?

d. Should measurement error be considered in setting a standard?

e. Should a standard of performance be set on each objective, on the total test score, or both?

f. When a standard for each objective measured by a test is set, how many objectives must an examinee pass to be identified as a "master"?

g. Should more than a single standard be set for each objective or on the total test?

h. Are there any subgroups of examinees for whom a different standard (or standards) should be set?

Use of Other Information

a. Should test score distributions of previous examinee groups and percent of masters information be used exclusively in setting standards (i.e., if 15% of the examinees failed last year, the same percent should fail this year), as background information for the judges, or not at all?

b. Should the test performance of contrasting groups be compared along with a consideration of the rates of false-positive and false-negative error rates with different standards?

c. Is there time in the procedure to collect valid criterion data (e.g., from "masters" and "non-masters") and test score performance for examinee groups of interest?

d. Should the scatter plot between scores on the test and an external criterion measure (e.g., a performance test) be studied?

e. Should a definition of "mastery" and "non-mastery" or the minimally competent candidate be prepared? If so, who should prepare it and how detailed should it be?

Data Analysis

a. When several groups of judges are used, how should differences in standards be resolved?

b. When more than one standard-setting method is used, how should differences be resolved?

c. Should weights be attached to different types of errors and used in the process of standard setting? If so, who should set the weights and how should it be done?